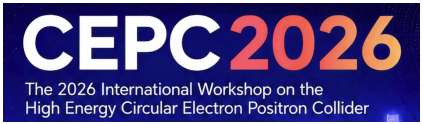
Towards precision frontier: development of reconstruction and analysis tools for the CEPC experiment

Marina Chadeeva for the LPI CEPC group

P.N. Lebedev Physical Institute of RAS, Moscow, Russia





Overview and outline

Parameters of interest in the "Higgs Factory" regime at 240 GeV

- branching fractions of Higgs boson decays, including invisible channels
- Higgs boson mass (sub-10 MeV precision expected [arXiv:2608.12071v1])

Searches for Higgs boson compositeness (if any) and CP structure, observation of rare decays, discrimination between ZZ and WW channels and many other tasks require unprecedented detector resolution and sophisticated analysis approaches.

CEPC experiment is expected to provide perfect tools to perform these studies.

In this talk preliminary results of four analyses are discussed.

Fast simulations with DELPHES

1. Correction for ISR in the analysis of recoil mass to $Z \to \mu\mu$
2. Novel approach to Higgs boson invariant mass reconstruction in 4-jet final states

Full simulations with CEPCSW

3. Correction for FSR in $Z \to ee$ to make resolution comparable with $Z \to \mu\mu$
4. Upper limit for invisible decays $H \to ZZ \to 4\nu$

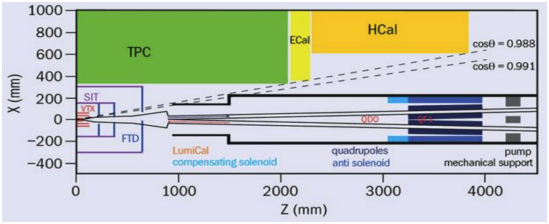
Work is supported by the RSF grant number 25-22-00716

1: Correction for ISR

Correction for ISR: simulation conditions

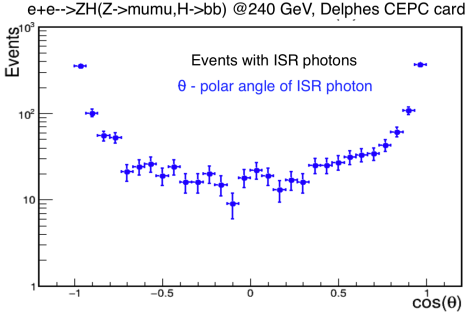
Simulation and reconstruction

- **Process**: $e^+e^- \to ZH$ with final state $Z \to \mu^+\mu^-$, $H \to b\bar{b}$ at $\sqrt{s}$ =240 GeV
- **Software**: WHIZARD3+Pythia6 with Initial State Radiation, DELPHES with CEPC card
- **Selection**:
  - two reconstructed isolated muons
  - reconstructed isolated photon in ECAL (up to $cos(\theta)$ =0.988) with E_γ >200 MeV



Efficiency

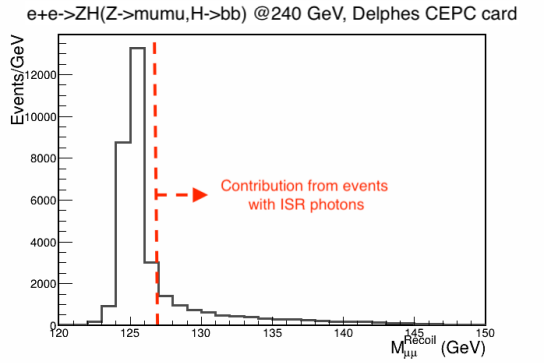
- 40000 generated events
- ∼85% events with two reconstructed muons
- ∼5% of events with two muons are accompanied by ISR photon reconstructed in ECAL

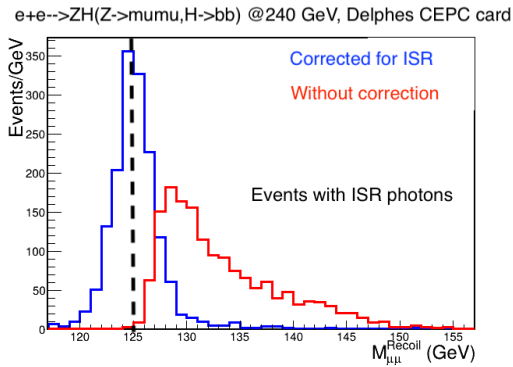


Correction for ISR: preliminary results

Distribution of recoil mass to $\mu\mu$-pair:
$M^{Recoil}_{\mu\mu} = \sqrt{(240 - E_{\mu\mu})^2 - (\vec{p_{1\mu}} + \vec{p_{2\mu}})^2}$

Correction by reco ISR photon: $M^{Recoil}_{\mu\mu\gamma} = \sqrt{(240 - E_{\mu\mu} - E_{\gamma})^2 - (\vec{p_{1\mu}} + \vec{p_{2\mu}} + \vec{p_{\gamma}})^2}$





Corrected distribution is wider than initial core but moves maximum to correct position; might be incorporated into simultaneous fits; additional room for correction – information from Lumical.

2: Analysis of 4-jet final states

Reconstruction of four-jet final states

Signal $e^+e^- \to Z(q\bar{q})H(b\bar{b}/gg/c\bar{c})$ ($\sigma \approx 95$ fb)

- About **50%** of final states from higgsstrahlung process
- Reconstruction issues
  - jet clustering
  - combinatorial (jet pairing)
  - jet energy resolution
- Backgrounds
  - leptonic: $e^+e^- \to l^+l^-$, $l^+l^-\nu_l\bar{\nu}_l$
  - semileptonic: $e^+e^- \to l^+l^-q\bar{q}$, $\nu\bar{\nu}q\bar{q}$
  - **Hadronic: $e^+e^- \to q\bar{q}$ ($\sigma \approx 54000$ fb), $WW \to 4j$ and $ZZ \to 4j$ ($\sigma \approx 7500$ fb), $Z(q\bar{q})H(\tau^+\tau^-, \tau \to hadr)$**
- Effective cut-based suppression for leptonic and sl
- ML with Transformer model for hadronic backgrounds suppression

Simulation and reconstruction

- **Process**: $e^+e^- \to ZH$ with final state $Z \to jj$, $H \to jj(b\bar{b}, c\bar{c}, gg)$ at $\sqrt{s}$ =240 GeV
- **Software**: WHIZARD1.9.5+Pythia6.4, DELPHES with CEPC_4th card, FastJet package
- **Samples**: 1 000 000 events per each dominating background (14), 100 000 events per each other background (15)
- $\mathcal{L}_{int} = 5.6$ ab^{-1}
- **Reconstruction**:
  - standard PFA out-of-the-box
  - exclusive ee-kt with FastJet (4 jets)
  - likelihood for jet pairing in Z and H
  - novel approach for boson mass reconstruction

Boson mass reconstruction in 4-jet final states

Dijet invariant mass

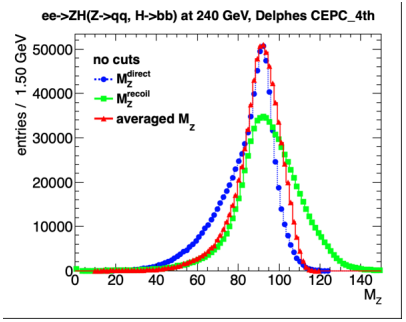
- For each pair of jets, invariant mass M_{ij} can be calculated in two ways ($i \neq j \neq k \neq l$):
  - **directly**: $M_{ij}^{direct} = \sqrt{E_{ij}^2 - (\vec{p}_{ij})^2}$
  - as **recoil** to another pair: $M_{ij}^{recoil} = \sqrt{(\sqrt{s} - E_{kl})^2 - (\vec{p}_{kl})^2}$
- Resulting mass is calculated as average: $\mathbf{M_{ij} = (M_{ij}^{direct} + M_{ij}^{recoil})/2}$

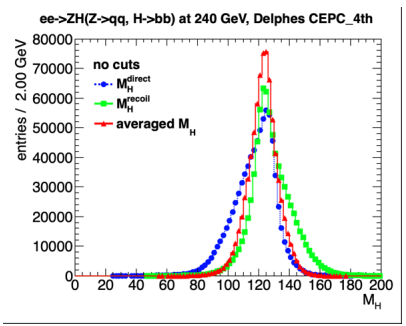
Z and H candidates

- calculate dijet masses for 6 combinations
- find best Z and H candidates using likelihood: $-ln(L_H) = \frac{(M_{ij}-m_Z)^2}{2\sigma_Z^2} + \frac{(M_{kl}-m_H)^2}{2\sigma_H^2}$

Advantages of proposed approach

- minimization of clustering issues
- resulting distribution is narrower

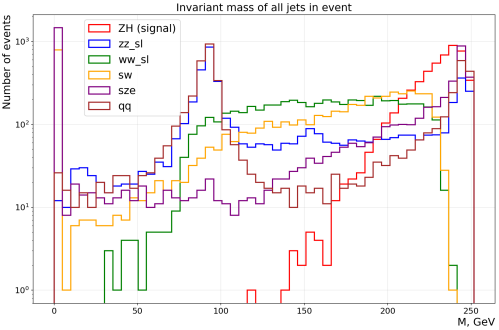


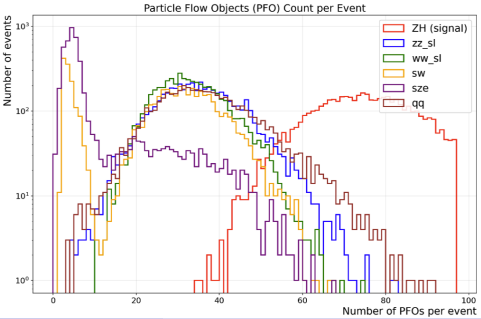


Background suppression: cut-based selections

Pre-processing and pre-selection

- **exactly 4 jets from ee-kt**
- **max likelihood** technique for jet association to boson candidates
- **set of kinematic cuts**:
  number of PFOs, mass windows, total reconstructed energy, jet directions, etc.
  (see backup for details)
- **effective leptonic and semileptonic background suppression – up to permille level after reweighting**
- **efficiency**:
  ∼60% of signal
  ∼1% of $q\bar{q}$ (still factor 10)
  ∼15-35% of diboson (still factor >20)

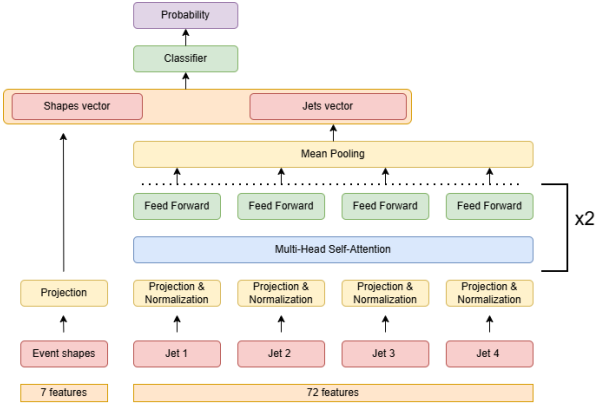




Hadronic background suppression

Transformer model

- **Input features per jet:**
  - Energy, $p_x, p_y, p_z, \phi, \cos\theta$
  - Number of neutral, charged PFOs
  - Max(D_{xy}, D_z signif.), D_{xy}, D_z, and next-to-max signif. of D_{xy}, D_z
  - Number of PFO with D_{xy}, D_z significance $> 3, > 5$
- **Input features per event:** thrust, sphericity, aplanarity, max. photon energy, likelihood of jet pairing, masses of H and Z candidates



Model parameters

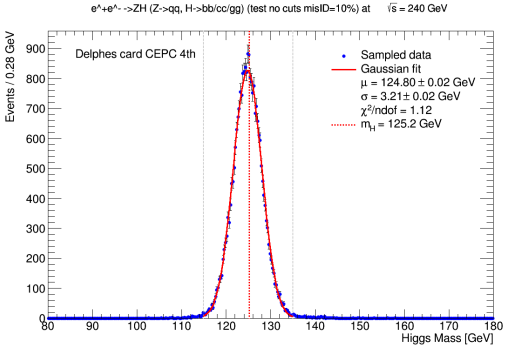
- **Architecture**: Self-Attention block + MLP classifier (256:64:32:1), loss: BCE
- **Samples**: signal (3×10^6) + 14 bgd (14×10^6), train:validation = 80:20; **realistic weights in test**
- Supervised learning, typical number of epochs ~40

Preliminary results: high-purity samples

Two types of trained classifiers: with and w/o pre-selections (cuts)

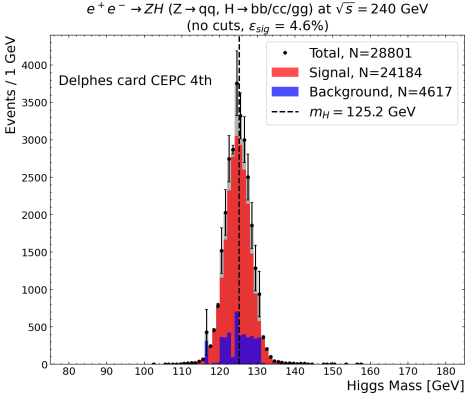
Signal shape after classifier

- pure signal, test sample
- Gaussian fits well (unbinned fit)
- central value biased, the same is before any selections, might be calibration

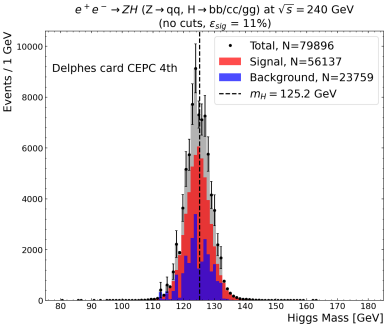


Signal + background

- low signal efficiency (4.6%)
- high purity (84%)
- bgd weights ⇒ large uncertainties

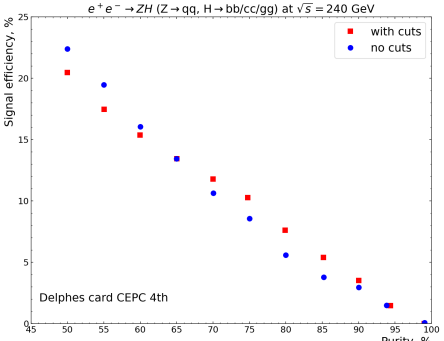


Preliminary results: low-purity samples



Efficiency vs purity

- training **with** and **w/o** pre-selections
- almost similar performance
- Gaussian signal shape



3: Correction for Final State Radiation

Impact of Final State Radiation

Motivation

- Precision of $Z \to ee$ reconstruction is worse than those of $Z \to \mu\mu$ channel, which results in twice worse accuracy, e.g.
  - of Higgs mass measurement [CPC 43, 4 (2019) 043002]
  - of setting upper limit for $H \to invisible$ [CPC 44, 12 (2020) 123001]
- One of the factors is **final state radiation** more intensive for electrons than for muons

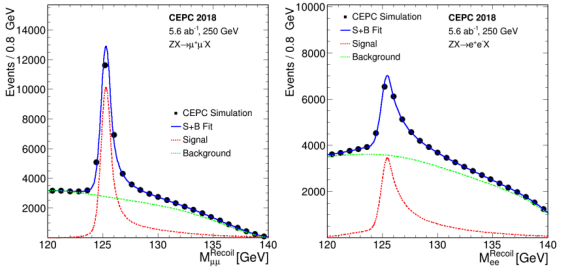


Illustration from [CPC 43, No. 4 (2019) 043002]

PF reconstruction

- Electron momentum from tracker, high accuracy ($\sim 2 \cdot 10^{-5}$ Gev^{-1} for 100 GeV)
- Photon energy from electromagnetic calorimeter, resolution $0.03 \div 0.2 \cdot \sqrt{E/GeV}$

FSR photons

- wide energy range (exponential)
- collinear with electrons ($cos(\alpha_{e\gamma}) > 0.99$)

Part 1: correction for FSR in ILD

ILD detector for ILC (model ILD_l5_v02)

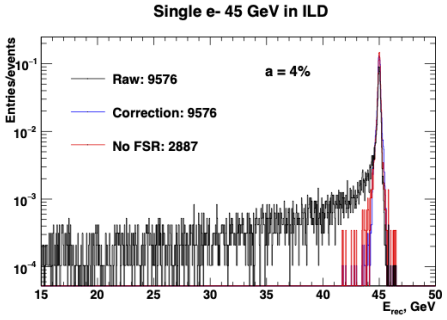
- tracker - TPC ($\sim 2 \cdot 10^{-5}$ GeV^{-1} for 100 GeV), 3.5 T magnetic field
- highly segmented ECAL ($\sim 0.17 \cdot \sqrt{E/GeV}$); allows "imitation" of smaller stoch. terms

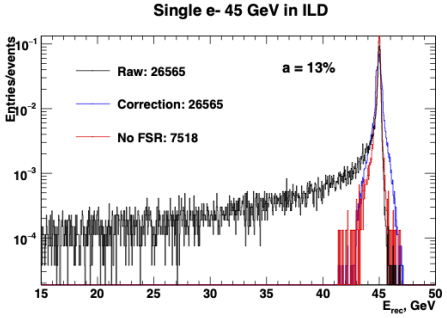
Full simulation and reco (ILCSoft v02-03-01)

- single electrons 10÷60 GeV (10 000 per energy)
- single gammas 0.2÷50 GeV for ECAL calibration

Correction

- 4-momenta of photons reconstructed within a cone $cos(\alpha_{e\gamma}) > 0.99$ are added to electron 4-momentum
- for stochastic $a_{ecal} = 0.04$ GeV$^{1/2}$ corrected distribution is similar to that w/o detected FSR
- RMS of corrected sample becomes about order of magnitude better even for $a_{ecal} = 0.13$ GeV$^{1/2}$





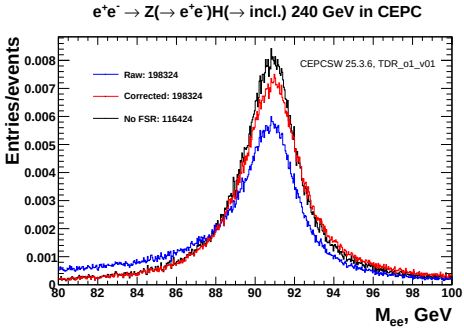
Part 2: correction for FSR in CEPC detector

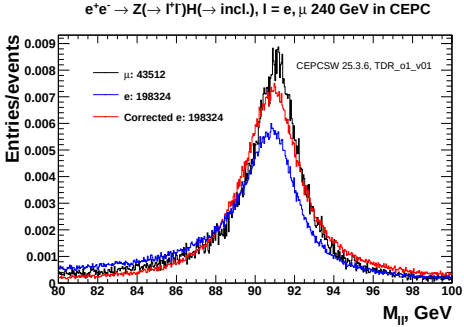
CEPC detector

- Detector model TDR_o1_v01
  - combined tracker ($\sim 2 \cdot 10^{-5}$ GeV^{-1} for 100 GeV)
  - crystal ECAL; expected $\leq 0.03 \cdot \sqrt{E/GeV}$
- $e^+e^- \to Z(l^+l^-)H(incl.), l = e, \mu$ at 240 GeV
- Full simulation and reconstruction
  - software CEPCSW 25.3.6, centrally processed reco
  - selection of two isolated leptons with $E_l > 15$ GeV
  - correction of electron energy by FSR photons within a cone $cos(\alpha_{e\gamma}) > 0.99$

Z boson mass reconstruction

- for $Z \to ee$, distribution with electron energies corrected by FSR is closer to that w/o detected FSR (for both electrons)
- distribution from $Z \to ee$ channel with corrected electron energies is very close to that obtained from $Z \to \mu\mu$ channel





4: Upper limit for $H \rightarrow$ invisible

Search for Higgs invisible decays

Status

- Standard model decay $H \to ZZ^* \to 4\nu$ with branching ~0.106%
- Current upper limit on branching ratio from ATLAS is 11%
- Studied in CEPC simulations for higgsstrahlung $e^+e^- \to ZH$ and final states $Z \to e^+e^-, \mu^+\mu^-, q\bar{q}$, $H \to invis$
  - CEPC publication : 0.26% for 5.6 ab^{-1} [CPC 44, 12 (2020) 123001] $\Longrightarrow$
  - CEPC TDR benchmark (2025): <0.05% for 20 ab^{-1}

Table 6. Expected precision of the measurement of BR($H \to$ inv) and the 95% CL upper limit on BR($H \to$ inv) for the dataset 5.6 ab^{-1}.

ZH final states	Precision of BR($H \to$ inv)×100 (%)	Upper limit on BR($H \to$ inv) (%)
$Z \to e^+e^-$, $H \to$ inv	45.37	1.08
$Z \to \mu^+\mu^-$, $H \to$ inv	23.57	0.55
$Z \to q\bar{q}$, $H \to$ inv	9.54	0.27
Combination	8.68	0.26

$Z \to q\bar{q}$ channel looks the most promising

Goal of this study (work in progress)

Investigate possible ways to analysis optimization for the final state $Z \to q\bar{q}$, $H \to invis$
Study different ML approaches

Search for Higgs invisible decays: cut-based approach

Production: higgsstrahlung $e^+e^- \to ZH$ at 240 GeV, $\mathcal{L}_{int} = 5.1$ ab^{-1}
Signal: final state $Z \to q\bar{q}$, $H \to inv$ (0.143 fb)

Dominating backgrounds

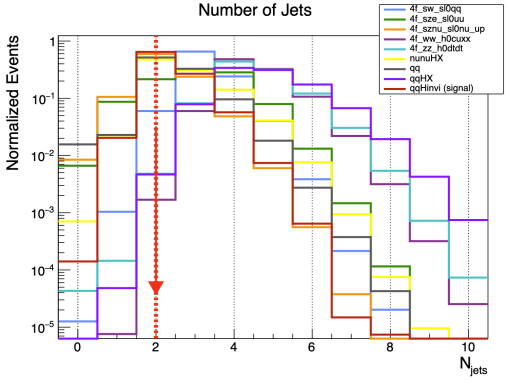
- $e^+e^- \to q\bar{q}$ ($\sim$50000 fb)
- $e^+e^- \to W^+W^-$ (hadr+sl, $\sim$14000 fb)
- $e^+e^- \to e\nu_e q\bar{q}$ (single W hadr, $\sim$3600 fb)
- $e^+e^- \to e^+e^-Z \to e^+e^-q\bar{q}$ (325 fb)
- $e^+e^- \to ZZ$ (hadr+sl, $\sim$650 fb)
- $e^+e^- \to \nu\bar{\nu}Z \to \nu\bar{\nu}q\bar{q}$ (147 fb)
- $e^+e^- \to q\bar{q}H$(incl.) (143 fb, except for signal)
- $e^+e^- \to \nu\bar{\nu}H$(incl.) (48 fb, except for signal)

Software

- CEPCSW 25.3.6, TDR_o1_v01
- FastJet, ee-kt inclusive ($R = 0.5$, $p = 1.0$)

Pre-selections

exactly 2 jets (rejects $\sim$30% of signal), no isolated leptons and photons, no photons with $E_\gamma >$30 GeV, $N_{pfo} \geq$6 in each jet



Cut-based background suppression

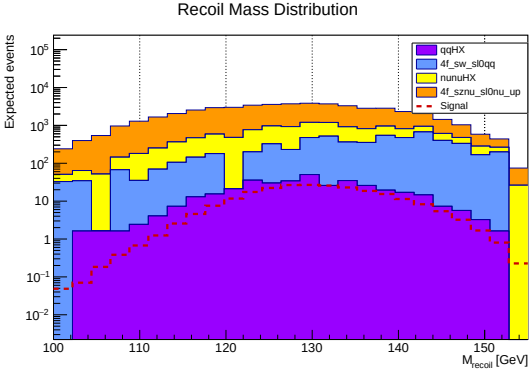
Pre-selection effectively rejects leptonic and semileptonic final states

Kinematic cuts

- azimuthal angle between two jets: $\Delta\phi_{jj} < 3.0$
- dijet transverse momentum: $20\ GeV < P_{Tjj} < 60\ GeV$
- dijet polar angle: $|\cos\theta_{jj}| < 0.7$
- dijet invariant mass: $75\ GeV < M_{jj} < 105\ GeV$
- missing 3-momentum ($\vec{P}_{miss} = -\sum_i^{Npfo} \vec{p_i}$): $20\ GeV < |\vec{P}_{miss}| < 80\ GeV$

Remaining after cuts and reweighting:

- **251 signal events** (∼35% of expected)
- **81754 background events**

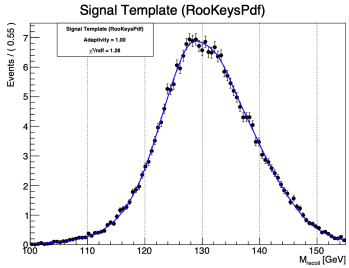


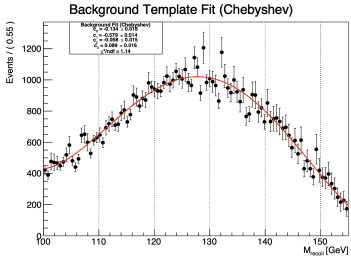
4 dominating background channels are shown

Likelihood ratio test for upper limit

Signal and background templates

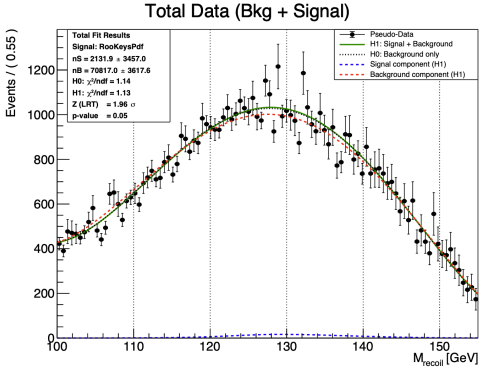
qqHinv sample of 135000 events





Analysed variable: M_{recoil} to dijet system

- complicated signal shape after selections - use KDE pdf
- background pdf: Chebyshev polynomial
- unbinned extended likelihood fit with varied signal strength μ, LR test against background only

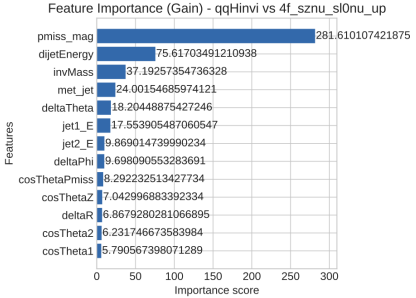


LR test: $\mu = 1.37$ at 95% CL $\Rightarrow$ UL $\approx 0.15\%$

Classifiers for dominating backgrounds (work in progress)

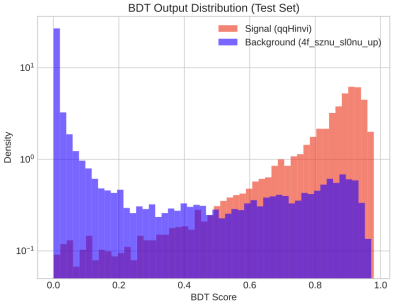
BDT from XGboost (after pre-selections)

- $\nu\bar{\nu}Z$ is the most severe (twice over all other bgd together), $\nu\bar{\nu}H$ is the next
- binary classifiers, supervised learning
- 13 input features: 5 kinematic cuts + 8 additional (jet parameters, jet distances, etc.)
- train:validation:test = 35:15:50
- signal:background = 1:1



Preliminary performance

Process	Cuts	BDT 1	BDT 2
$q\bar{q}H(H \to inv)$	54%	81%	83%
$\nu\bar{\nu}Z$	15%	12%	–
$\nu\bar{\nu}H$	14%	–	7%



No significant gain so far for $\nu\bar{\nu}Z$...

Summary and plans

Two studies have been performed of higgsstrahlung process at 240 GeV in CEPC using fast simulation with DELPHES package.

Correction for initial state radiation

- Higgs boson mass reconstructed as recoil to $\mu\mu$ system in the process $e^+e^- \to ZH$ ($Z \to \mu^+\mu^-$, $H \to b\bar{b}$) can be corrected for ISR in about 5% of events using photon detected in the ECAL.
- Plans are to check this with full simulation.

Studies of 4-jet final states

- Higgs boson mass reconstruction has been studied in 4-jet final states with hadronic decays $H \to b\bar{b}/c\bar{c}/gg$.
- Novel approach to boson mass reconstruction is proposed, which helps to get near Gaussian shape of reconstructed signal.
- Transformer-based classifier was trained to reject hadronic backgrounds, achieved performance in terms of signal efficiency/purity is: 20%/50%, 10%/70%, 4.5%/84%.
- Investigations with full simulation are planned including study of systematic uncertainties

Summary and plans (cntd.)

Two studies have been performed for higgsstrahlung process at 240 GeV in CEPC using full simulations.

Correction for final state radiation

- Correction for FSR has been studied on single particles using ILD simulations with ILCSoft for different ECAL resolutions.
- In the CEPC simulations, the correction for FSR was demonstrated to improve Z boson mass resolution from $Z \to ee$ decay.
- Cross checks are planned of lepton reconstruction efficiencies.
- Impact of correction on physics analysis will be investigated.

Setting upper limit on $H \to invisible$ branching

- Invisible Higgs boson decay $H \to ZZ^* \to 4\nu$ in the process $e^+e^- \to ZH(Z \to q\bar{q}, H \to invisible)$ has been studied. With cut-based approach to background suppression, upper limit was estimated to be 0.15% at 95% CL (w/o systematic uncertainties).
- Development of ML-based classifier is in progress.

We would like to thank our IHEP colleagues for providing generated and reconstructed samples and CEPC software!

Backup slides

Cuts for 4-jet final states selections

Pre-selections

- Total $N_{PFO} \geq 40$
- Charged $N_{PFO} \geq 25$
- Min($N_{PFOinJet}$) ≥ 6
- 2-nd min($N_{PFOinJet}$) ≥ 10
- Max(D_{xy}, D_z significance)≥ 5
- Second max(D_{xy}, D_z significance)≥ 5
- Aplanarity ≥ 0.06
- Total energy ≥ 170 GeV
- Max(iso photon energy) ≤ 40 GeV
- Min(jet energy) ≥ 15 GeV
- Total invariant mass ≥ 150 GeV
- $m_Z \in [55; 110]$ GeV (Z candidate)
- $m_H \in [90; 150]$ GeV (H candidate)
- $(m_H - m_Z) \in [-20; 65]$ GeV
- $cos(\theta_{sum}) \leq 0.9$

